The Promise and Pitfalls of Data Mining: Ethical Issues

William Seltzer
Fordham University
(seltzer@fordham.edu)

Abstract

The paper reviews three major ethical issues that arise in data mining, particularly data mining involving one or more federal statistical data sets, in terms of the ASA's Ethical Guidelines. They are the suitability and validity of methods used, privacy and confidentiality, and the objectives of the data mining effort. The paper also identifies three imperatives that emerge from this review: (a) the need to switch attention from disclosure risk to disclosure harm, (b) the special duties owed to vulnerable individuals and population subgroups, and (c) the ethical responsibilities of statistical agency leadership and senior staff.

Keywords: federal statistical system, harm, risk of disclosure, statistical confidentiality, targeting

I. Introduction

These comments are in large part drawn from a presentation I made at a November 2004 ASA-NISS Conference on Statistics and Counterterrorism[1] and a paper, "Frequently Asked Questions Regarding the Privacy Implications of Data Mining" now being prepared on behalf of the ASA Committees on Privacy and Confidentiality, Professional Ethics, and Scientific and Public Affairs.

My comments are also informed by my experience as the chair of the ASA Committee on Professional Ethics and related research I and others have carried out on the use population data systems to target individuals and vulnerable population sub-groups for human rights abuses. However, I would emphasize that the views expressed here are my own and do not necessarily reflect those of the ASA or its Committee on Professional Ethics.

Like most statistical methodologies data mining by itself is ethically neutral. This is particularly so because the term data mining is a generic one referring to a wide range of procedures, involving diverse data sets, and carried out for numerous purposes. For these reasons there are no specific references to data mining in the ASA's *Ethical Guidelines for Statistical Practice*, adopted by the ASA Board in 1999 and available on line at the ASA's website (www.amstat.org) and in print from the ASA office. Moreover, it should be understood that, whether one is dealing data mining or some other topic, ethical standards and legal requirements are not the same thing. In other words, while there is a very large overlap between the unlawful and the unethical, the two concepts are not equivalent.

From the perspective of statistical practice, data mining raises three quite different sorts of ethical issues. These are: (a) the suitability and validity of the methods employed in any given data mining application, (b) the degree to which confidentiality and privacy obligations are respected, and (c) the over-all aims of a given data mining application. Each of these general issues are addressed in the ASA's *Ethical Guideline for Statistical Practice.*

II. Suitability and Validity

Several provisions of the ASA's ethics guidelines address issues of the suitability and validity of methods used in any statistical application, including data mining. They include, in section II.A,

> 2. Guard against the possibility that a predisposition by investigators or data providers might predetermine the analytic result. Employ data selection or sampling methods and analytic approaches that are designed to assure valid analyses in either frequentist or Bayesian approaches.
>
> 4. Assure that adequate statistical and subject-matter expertise are both applied to any planned study. If this criterion is not met initially, it is

[1] Seltzer, W. (2005), "Statistics and Counterterrorism: The Role of Law, Policy and Ethics," 2005 Proceedings of the American Statistical Association, Section on Risk Analysis [CD-ROM], Alexandria, VA: American Statistical Association.

> important to add the missing expertise before completing the study design.
>
> 5. Use only statistical methodologies suitable to the data and to obtaining valid results. For example, address the multiple potentially confounding factors in observational studies, and use due caution in drawing causal inferences.
>
> 7. The fact that a procedure is automated does not ensure its correctness or appropriateness; it is also necessary to understand the theory, the data, and the methods used in each statistical study. This goal is served best when a competent statistical practitioner is included early in the research design, preferably in the planning stage.

And, in section II.C,

> 2. Report statistical and substantive assumptions made in the study.
>
> 5. Account for all data considered in a study and explain the sample(s) actually used.
>
> 6. Report the sources and assessed adequacy of the data.
>
> 7. Report the data cleaning and screening procedures used, including any imputation.
>
> 8. Clearly and fully report the steps taken to guard validity. Address the suitability of the analytic methods and their inherent assumptions relative to the circumstances of the specific study. Identify the computer routines used to implement the analytic methods.
>
> 9. Where appropriate, address potential confounding variables not included in the study.
>
> 12. Report the limits of statistical inference of the study and possible sources of error. For example, disclose any significant failure to follow through fully on an agreed sampling or analytic plan and explain any resulting adverse consequences.

III. Privacy and Confidentiality

The ASA ethics guidelines address privacy and confidentiality obligations in section II.D, "Responsibilities to Research Subjects (including census or survey respondents and persons and organizations supplying data from administrative records, as well as subjects of physically or psychologically invasive research)." Among the pertinent provisions are

> 1. Know about and adhere to appropriate rules for the protection of human subjects, including particularly vulnerable or other special populations who may be subject to special risks or who may not be fully able to protect their own interests ... Laws of other countries and their subdivisions and ethical principles of other professional organizations may provide other guidance.
>
> 3. Avoid excessive risk to research subjects and excessive imposition on their time and privacy.
>
> 4. Protect the privacy and confidentiality of research subjects and data concerning them, whether obtained directly from the subjects, from other persons, or from administrative records. Anticipate secondary and indirect uses of the data when obtaining approvals from research subjects; obtain approvals appropriate for peer review and for independent replication of analyses.

IV. The Aims of a Data Mining Effort

Finally, when considering the over-all aims of any data mining application, two provisions of the Preamble to the ASA ethics guidelines seem particularly pertinent. The first is from a section on "Statistics and Society,"

> Statistical tools and methods, like many other technologies, can be employed either for social good or for evil. The professionalism encouraged by these guidelines is predicated on their use in socially responsible pursuits by morally responsible societies, governments, and employers. Where the end purpose of a statistical application is itself morally reprehensible, statistical

> professionalism ceases to have ethical worth.

The second from a section "Shared Values,"

> All statistical practitioners are obliged to conduct their professional activities with responsible attention to:
>
> 1. The social value of their work and the consequences of how well or poorly it is performed. This includes respect for the life, liberty, dignity, and property of other people.

Other professional associations in statistics and allied fields also contain guidance applicable to data mining. Given the central role that data processing technology plays in data mining, the Association for Computing Machinery's 1992 "Code of Ethics and Professional Conduct," available at http://www.acm.org/constitution/code.html provides particularly relevant guidance. For example, two of its "moral imperatives," state

> 1.1 Contribute to society and human well-being. -- This principle concerning the quality of life of all people affirms an obligation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health and safety. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and will avoid harmful effects to health and welfare.
>
> 1.2 Avoid harm to others. -- ... This principle prohibits use of computing technology in ways that result in harm to any of the following: users, the general public, employees, employers ... Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event the responsible person or persons are obligated to undo or mitigate the negative consequences as much as possible. One way to avoid unintentional harm is to carefully consider potential impacts on all those affected by decisions made during design and implementation.

Two additional sets of guideline relating to statistics exist at the international level, the International Statistical Institute's *1985 Declaration on Professional Ethics*, which is available at: http://www.cbs.nl/isi/ethics.htm and the United Nations Statistical Commission's 1994 Fundamental Principles of Official Statistics, which is available at: http://unstats.un.org/unsd/goodprac/bpabout.asp. Although, like the other available guidelines, neither of these sources directly discuss data mining, the advice provided roughly parallels that contained in the ASA guidelines.

V. Some Concluding Thoughts

What are some of the main imperatives that emerge from this brief review of ethical norms relating to statistical applications involving data mining, particularly where one or more of the data sets used was generated by a federal statistical agency?

First is the need to clarify what we mean by "risk." (Typically, the term risk arises when considering the risks-benefits trade offs associated with a given data mining application.) Second, we need to give special attention to protecting, in the language of section II.D.1 of the ASA guidelines, "vulnerable or other special populations who may be subject to special risks or who may not be fully able to protect their own interests." Third, the leadership of federal statistical agencies and their senior staffs have to more clearly recognize, and then better internalize, an awareness of the ethical dimensions of their responsibilities.

Before commenting on each of these areas in turn, it is essential to remind ourselves about an unfortunate body of experience that touches on each of these issues. I refer to the role that population data systems in the United States and elsewhere have on occasion played in targeting vulnerable individuals or population subgroups for investigation, prosecution, forced migration, or more serious human rights abuses. Links to several papers, some with extensive bibliographies, describing the use of official statistics to target individuals and members of vulnerable population subgroups, and to related issues of statistical confidentiality, may be found at http://www.uwm.edu/~margo/govstat/integrity.htm Clearly, this experience needs to be taken into account in any discussion of risks, vulnerable population subgroups, and the ethical responsibilities of statisticians.

At present nearly all discussions among statisticians about the risks associated with data mining or other methods of data dissemination and analysis focus on the risk of disclosure, that is, the risk that a respondent in a given data set can be identified. This is not surprising, since this is a concept that often can be studied directly in quantitative terms through statistical analysis. However, given the extreme consequences that have sometimes flowed from disclosures based on population data systems in the past a far more pertinent risk is the risk that a particular method of data dissemination or analysis, including data mining, can result in substantial harm to the individual or the concerned population subgroup. Thus, it is the nature and consequences of a disclosure, rather than its relative frequency, that emerges as the major item of concern.

Let me turn now to the widely recognized ethical principle, also embodied in the ASA guidelines as discussed above with respect to statistical confidentiality, that we owe a special duty toward vulnerable populations and others who may not be fully able to protect their own interests. Many will recall the problems encountered by Senator Kennedy of Massachusetts in repeatedly being denied permission to board commercial airlines because his name was on a Transportation Security Administration "do not fly" list because of concerns about someone else with the same name. Not surprisingly, Senator Kennedy was, with some effort and time, able to resolve the matter. But if you are not Senator Kennedy or some other prominent person with the resources and contacts to get redress, the task of dealing with a "false match" may not be a simple one. The potential threats to members of vulnerable population subgroups (for example, people who speak little or no English, recent immigrants, Arab Americans generally and those mistaken for them, persons poorly educated or living in poverty regardless of origin) is even greater.

Finally, in considering the ethical responsibilities of statisticians in federal statistical agencies, particularly senior statisticians and the agency leadership, two lessons seem evident.

First, the responding public expects a statistical agency and its leadership to behave ethically. In the words of Census Director Kincannon, commenting on criticism the Bureau received following the Bureau's provision of 5digit zip code data from the 2000 Census on persons of Arab American ancestry to the Department of Homeland Security (DHS), "We recognize that simply making sure we obey the law may not always be enough to ensure that people trust us ...Perception also affects how people view and cooperate with the census" (*New York Times*, 8/31/2004, p. 14). Commenting on the same event, his predecessor as Census Director, Kenneth Prewitt stated, "There is an issue of principle involved as well as law" (*New York Times* (7/29/2004, p. 19). This comes back to a point made at the outset of this paper, "ethical standards and legal requirements are not the same thing ... while there is a very large overlap between the unlawful and the unethical, the two concepts are not equivalent."

Second, the responding public can have very long memories when trust is betrayed and the outputs of the national statistical system are misused to target vulnerable population subgroups. For example, in the Netherlands, where, in the early 1940s, data derived from its population registration system was used in the round up and deportation of Jews and Roma to death camps during the Holocaust, Statistics Netherlands today has one of the lowest survey response rates in Europe and was forced to abandon its traditional population census in the late 1970's because of issues of respondent mistrust.

Similarly, in this country, when the Bureau's special assistance to DHS in providing tabulations and a data file from the 2000 Census on persons of Arab American ancestry became widely known, Arab American leaders, former Census Director Prewitt, and representatives of other minority groups on the Bureau's decennial census advisory committee directly referred to the Bureau's involvement in the round up of the Japanese American population on the West Coast in the months after Pearl Harbor. Reflecting on this experience, former Director Prewitt, remarked, "In World War II we violated our principles even if we didn't violate the law, and we assured people we wouldn't do it again" (*New York Times* (7/29/2004, p. 19).

We assume that the leaders in most statistical agencies are moral persons, striving to do the right thing. However, in the set of competing priorities (such as, law, science, user needs, budgetary constraints politics) they seek to reconcile, ethics are rarely considered explicitly. As argued here, public expectations and the public's long memory provide two strong utilitarian motives for explicitly taking ethical considerations into account.

As stated at the outset, data mining is ethically neutral. What we do with it, however, is not. Since some of the biggest ethical tragedies have occurred when people either did not recognize there was an ethical issue involved or were unable to take part in a discussion about it, it is important to consider and exchange views on the ethical issues associated with data mining efforts. This paper is simply a contribution to that discussion.